Journal of Biomedical Informatics 43 (2010) 397–406

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers

Rosalia Maglietta [a], Angela Distaso [a], Ada Piepoli [b], Orazio Palumbo [c], Massimo Carella [c], Annarita D'Addabbo [a], Sayan Mukherjee [d], Nicola Ancona [a,*]

[a] Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Via Amendola 122/D-I, Bari, Italy
[b] Unità Operativa di Gastroenterologia, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale, 71013 San Giovanni Rotondo (FG), Italy
[c] Servizio di Genetica Medica, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale, 71013 San Giovanni Rotondo (FG), Italy
[d] Institute for Genome Science and Policy, Duke University, Durham, NC, USA

ARTICLE INFO

Article history:
Received 26 June 2009
Available online 29 September 2009

Keywords:
Bioinformatics
Microarray data
Pathway
Colorectal cancer

ABSTRACT

One of the major problems in genomics and medicine is the identification of gene networks and pathways deregulated in complex and polygenic diseases, like cancer. In this paper, we address the problem of assessing the variability of results of pathways analysis identified in different and independent genome wide expression studies, in which the same phenotypic conditions are assayed. To this end, we assessed the deregulation of 1891 curated gene sets in four independent gene expression data sets of subjects affected by colorectal cancer (CRC). In this comparison we used two well-founded statistical models for evaluating deregulation of gene networks. We found that the results of pathway analysis in expression studies are highly reproducible. Our study revealed 53 pathways identified by the two methods in all the four data sets analyzed with high statistical significance and strong biological relevance with the pathology examined. This set of pathways associated to single markers as well as to whole biological processes altered constitutes a signature of the disease which sheds light on the genetics bases of CRC.

1. Introduction

Identifying individual genes and gene networks involved in onset and progression of complex and polygenic diseases is a major challenge of the current research in medicine and, in particular, in oncology [1,2]. The modern DNA microarray technologies play a fundamental role in the achievement of this ambitious objective as they allow to obtain quantitative, genome-wide descriptions of the expression levels of genes in tissues under different phenotypic conditions [3].

Although the great potential offered by these technologies, there is a unanimous consensus to judge with caution the results obtained by DNA microarray experiments because they are poorly reproducible. In fact, when we compare results obtained by different microarray studies which examine the same biological conditions, e.g. differential expression between tumor and normal samples, the lists of differentially expressed (DE) genes show little overlap [4,5]. Moreover, dissimilar lists of DE genes also result when different statistical approaches are used for analyzing the data [6]. In [4], for example, the authors analyzed three public data sets, consisting of normal and tumor samples of prostate, by using two distinct statistical methods: Statistical Analysis of Microarray (SAM) [7] and Mixed Model Analysis (MMA) [8]. Both methods produced lists of DE genes having only the 6% of DE genes common in the three data sets. Moreover, the overlap reduced to 3% considering the common DE genes identified by the two methods.

The reasons of this lack of reproducibility stem from (a) the univariate statistics adopted which do not take into account gene interactions and (b) the need of limiting the effects of multiple hypothesis testing.

To overcome these shortcomings which make results by microarray experiments difficult to compare, a new trend has emerged recently in computational biology in which the activity of a gene or of a whole biological process in a disease is assessed by using sets of genes [9–13]. These gene sets code biological pathways, such as cellular functions and biological processes, or represent a unique signature of deregulation of a given gene [14]. The former are manually, knowledge-driven built gene sets in which annotated genes are grouped on the basis of evidences coming from knowledge bases and literature. The latter are experimentally derived by analyzing the cell response to a given variation. In [11], for example, the pathway or signature associated to the activity of a given oncogene is defined as the set composed of those genes that most differentially express under the perturbation (impulse) of the oncogene. So, for evaluating the activity of a given gene in

* Corresponding author.
E-mail address: ancona@ba.issia.cnr.it (N. Ancona).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.09.005

a disorder we could measure the amount of deregulation or enrichment of its signature in the given experimental conditions. This approach is analogous to procedures developed in the general framework of system theory in which the properties of a given system are studied characterizing the response of the system to the impulse [15]. Under this perspective the problem of measuring gene deregulation is an inverse problem because we want to detect an event measuring its effects. Inverse problems are in general ill posed and in this particular case the size of the signature acts as regularization parameter [16].

In this paper, we address the problem of assessing the reproducibility of results of pathway analysis obtained by microarray experiments. In particular, we assess whether lists of pathways found associated to given phenotypic conditions, determined in independent microarray experiments, share biologically relevant and statistically significant pathways. To this end, we analyzed data sets relative to four independent genome-wide expression studies of colorectal cancer which compared normal versus tumor tissues. Three data sets come from literature [17–19] and the fourth is a new data set collected in Casa Sollievo della Sofferenza Hospital, Foggia, Italy. In our analysis we used Molecular Signatures Database (MSigDB) [9], a vast collection composed of 1891 curated gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. Moreover, to evaluate how the adopted statistical model influences the reproducibility of the results, we applied two well-known and statistically well-founded approaches which assess pathway enrichment in expression studies: Gene Set Enrichment Analysis (GSEA) [9] and Random-Set Methods (RS) [20]. Although these approaches aim at evaluating the association of pathways with phenotypes they are deeply different. Moreover, they use different strategies for assessing the statistical significance of the deregulation of pathways in the experimental conditions analyzed.

Our study shows that pathway based approaches are suitable for dissecting complex and polygenic diseases and provide significant and highly reproducible results. In particular, our analysis highlights a signature of colorectal cancer composed of 53 statistically significant and biologically relevant pathways found deregulated in all the four data sets by both methods. The biological relevance of this set of pathways in the pathology is analyzed in depth. Finally, we provide a set of suggestions to users of gene set methods.

2. Materials and methods

2.1. Data set description

Four microarray gene expression data sets of tumor vs normal human colon tissues were analyzed in our study. The first data set (SGR1) is composed of 22 normal and 25 tumor specimens, profiled by using the Affymetrix HGU133A GeneChip (22283 probesets) [17]. The microarray data are accessible through ArrayExpress site, with Accession No. E-MTAB-57. The second data set (JIANG) is composed of 24 pairs of normal and tumor colon specimens profiled by using Illumina BeadChip Human Ref8-v2 (22184 probesets) [19]. The data can be downloaded by Gene Expression Omnibus with Accession No. GSE10950. The third data set (GARD) is composed of 20 paired tumor-normal colon samples, profiled by using the Affymetrix GeneChip Human Exon Array 1.0 ST (22011 genes) [18]. Finally, we analyzed a new data set (SGR2) collected in Casa Sollievo della Sofferenza Hospital, Foggia, Italy, composed of 14 paired tumor-normal samples, profiled by using the Affymetrix GeneChip Human Exon 1.0 ST (22011 genes). The data sets were normalized by using the Robust Multi-array Average (RMA) procedure [22,40].

2.2. Gene sets

The database of gene sets used in this work was the C2 collection of the Molecular Signatures Database (MSigDB) [9]. This collection consists of 1891 curated gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts.

2.3. Algorithms

We are given a data set $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ composed of ℓ labeled specimens, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ for $i = 1, 2, \ldots, \ell$ and d is the number of probes on the microarray in the adopted technology. Let us suppose we have ℓ_+ positive and ℓ_- negative examples, such that $\ell = \ell_+ + \ell_-$. Moreover, we are given a gene set $G = \{g_1, g_2, \ldots, g_m\}$ composed of m probes, where $m \ll d$.

2.3.1. RS

Let s_i, $i = 1, \ldots, d$, be a score associated to each probe. This score is a quantitative measure of differential expression which in our case is based on a two sample t-statistic for each gene t_i, the two samples are the two phenotypes or conditions. Specifically, $s_i = \Phi^{-1}(\widehat{F}(t_i))$, $i = 1, \ldots, d$, where t_i were the two-sample t-statistics values computed for each gene, $\widehat{F}(t_i) = rank(t_i)/d$ where $rank(t_i)$ is the rank of the value t_i in the array $[t_1, \ldots, t_d]$, and Φ is the standard normal cumulative distribution function. Given these scores the measure of gene set deregulation is $Z = (\overline{X} - \mu)/\sigma$, where $\overline{X}$ is the average of gene scores, $\overline{X} = \frac{1}{m}\sum_{g \in G} s_g$, and $\mu = \mathbb{E}\{\overline{X}\}$ and $\sigma = \mathrm{var}\{\overline{X}\}$ are easily computed from the full set of gene scores. Large positive or negative values of Z are expected if G is up-regulated or down-regulated in the experimental conditions analyzed. P-values are computed using a non-parametric permutation test [23] with 1000 permutations of the phenotypic labels and false discovery rate (FDR) computations are provided using the method described in [24].

2.3.2. GSEA

This method uses a variation of a Kolmogorov–Smirnov statistic to provide an enrichment score for each gene set [9]. As in the random set method a score measuring the correlation of a probe with the phenotype is required, s_i, $i = 1, \ldots, d$. We use the signal-to-noise metric in the standard GSEA setting as our score. This metric is very similar to the two sample t-statistic used in our implementation of RS. Based on these correlation scores and the adjusted Kolmogorov–Smirnov statistic we compute an enrichment score which is signed. Negative scores correspond to down-regulation of the gene set and positive scores correspond to up-regulation of the gene set. These enrichment scores are then normalized to take into account the size of the gene sets resulting in a normalized enrichment score. This normalization is done based on phenotypic permutations followed by standardization, see [9]. P-values as well as false discovery rates are computed using the standard setting of the software.

3. Results

We evaluated the deregulation of all the gene sets belonging to MSigDB in the four microarray gene expression data sets independently. In particular, for having more precise insights into the cellular mechanisms involved in the pathology at hand, we assessed up-regulated and down-regulated gene sets separately. With the term up(down)-regulated pathway we mean a gene set enriched of genes up(down)-regulated with respect to a given phenotypic condition. Up- and down-regulation was always referred to tumor

phenotype. The level used in each test to assess deregulation was $\alpha = 0.05$. Table 1 reports the number of up/down-regulated pathways identified by the two methods in each data set and the estimated FDR. For example, in SGR1 data set, GSEA identified 240 up- and 203 down-regulated gene sets with $P \leqslant 0.05$ and FDR = 32% in both cases. In the same data set, RS identified 166 up- and 112 down-regulated gene sets with $P \leqslant 0.05$ and FDR = 31%. The number of common gene sets identified by both methods was 155 (up) and 94 (down) as reported in the last two columns of Table 1. The number of common gene sets identified by the two methods was statistically significant as assessed by using the Fisher's exact test. A complete description of the results obtained by the two methods on the four data sets is given as supplemental material.

As Table 1 shows, the up-regulation signals are stronger than the down-regulation ones, consistently in all the data sets analyzed. In fact, the range of up-regulated pathways is [148, 356] and [164, 228] as identified by GSEA and RS, respectively. The range of down-regulated pathways, on the contrary, is [23, 298] for GSEA and [57, 137] for RS, and in two data sets (SGR2 and GARD) the number of down-regulated gene sets is smaller than the expected number of pathways deregulated by chance. In fact, in the case of independency between gene expression levels and phenotype, we expect to find $n\alpha = 95$ pathways deregulated at 0.05 level by chance, where $n = 1891$ is the number of assessed gene sets.

This consideration is confirmed by analyzing (a) the number of pathways up- and down-regulated identified simultaneously by the two methods in each data set (the last two columns of Table 1) and (b) the number the pathways which are consistently and significantly up- or down-regulated in all the data sets (the last row of Table 1). In fact, the number of gene sets identified simultaneously by the two methods ranges in [98, 189] (up) and [19, 129] (down). Moreover, the number of pathways up-regulated in all the data sets with $P \leqslant 0.05$ is 73 identified by GSEA and 87 identified by RS. Both methods identified only five gene sets down-regulated in all the data sets with $P \leqslant 0.05$.

Intersecting the lists of gene sets identified independently by the two methods, with $P \leqslant 0.05$, in all the data sets analyzed, we determined 52 up- and 1 down-regulated pathways. In Tables 5 and 6 we show the name of the 53 gene sets identified by both methods together with the statistical parameters relative to each gene set as evaluated by GSEA and RS, respectively. The last column reports the median rank of the pathway in the four data sets. This list of 53 gene sets constitute a signature of the pathology in terms of pathways deregulated and we use it as gold standard for our successive statistical assessments. In supplemental materials, we report the statistical parameters relative to this set of pathways as estimated by the two methods on the four data sets. Moreover, we provide a detailed analysis concerning the biological and functional relevance of these pathways in colorectal cancer at the end of this section.

After having defined the gold standard, i.e. the pool of pathways significantly altered in the phenotypic conditions examined, it is interesting to estimate the rate of false positive (FP) pathways identified in an experiment as a function of (a) the statistical method adopted for assessing deregulation and (b) the available data. In Table 2, we report the FP rate evaluated comparing the set of pathways associated to the phenotype with $P \leqslant 0.05$ and the gold standard. Although RS performs slightly better than GSEA in terms of FP pathways, the rate of FP gene sets identified as deregulated by analyzing one data set only is extremely high. A possible strategy for reducing the FP rate is intersecting lists of deregulated pathways identified by using different data sets. In Table 3, we report the number of pathways simultaneously deregulated in two data sets and the corresponding FP rate. The rate of FP up-regulated pathways reduces to a median value of 57% for both methods and this value reduces to 46% if we intersect lists obtained by analyzing three different data sets (see Table 4).

Table 2
Percentage of false positive pathways identified by the two methods in the four data sets.

	GSEA		RS	
	UP (%)	DOWN (%)	UP (%)	DOWN (%)
SGR1	78	99	69	99
SGR2	65	98	68	98
JIANG	85	99	73	99
GARD	76	96	77	98

The last consideration concerns the methods adopted for assessing deregulation. Although their statistical bases are deeply different, our analysis shows that they perform similarly in all the experimental conditions analyzed. In particular the lists of pathways associated to the phenotype produced by the two methods show a significant overlap. We measured the intersection of the two rank-ordered gene set lists produced by GSEA and RS as a function of the number of considered gene sets (see Fig. 1). As the picture shows, intersecting the lists composed of the mostly deregulated 150 gene sets produced by each method, the overlap range from 62% to 77% in the four data sets.

4. Biological discussion

We analyzed in depth the biological relevance of the 53 pathways identified as deregulated in the four data sets by the two methods, listed in Table 5 or 6. We reported a brief description of the single pathways belonging to this group, which could be used as diagnostic biomarkers, and underlined their importance in oncogenesis. Most gene sets have been already shown to be involved in colorectal tumorigenesis. Some of these pathways are related to cell cycle, whose deregulation has been identified as one of the hallmarks of cancer [32]. Numerous genes that change expression during colon cancer progression encode proteins related to cell cycle and proteins involved in growth and differentiation [28]. The human cell cycle in normal somatic cells is characterized by its high precision. This remarkable accuracy is achieved by a number of signal transduction pathways, known as checkpoints,

Table 1
Number of up- and down-regulated pathways identified by the two methods with $P \leqslant 0.05$ and the corresponding FDR.

	GSEA				RS				Overlap	
	UP	FDR (%)	DOWN	FDR (%)	UP	FDR (%)	DOWN	FDR (%)	UP	DOWN
SGR1	240	32	203	32	166	31	112	31	155	94
SGR2	148	100	46	100	164	39	57	39	98	23
JIANG	356	23	298	32	190	27	137	27	189	129
GARD	217	81	23	89	228	29	61	29	155	19
Overlap	73		5		87		5		52	1

Table 3
Number of pathways identified by (a) GSEA and (b) RS in two data sets simultaneously. The FP rate is reported in parentheses. The upper and lower triangular parts of the tables contain the number up- and down-regulated pathways, respectively.

	SGR1	SGR2	JIANG	GARD
(a)				
SGR1	□	103 (50%)	204 (75%)	128 (59%)
SGR2	29 (97%)	□	112 (54%)	94 (45%)
JIANG	102 (99%)	25 (96%)	□	170 (69%)
GARD	14 (93%)	9 (89%)	11 (91%)	□
(b)				
SGR1	□	103 (50%)	128 (59%)	120 (57%)
SGR2	25 (96%)	□	114 (54%)	123 (58%)
JIANG	54 (98%)	20 (95%)	□	150 (65%)
GARD	21 (95%)	16 (94%)	16 (94%)	□

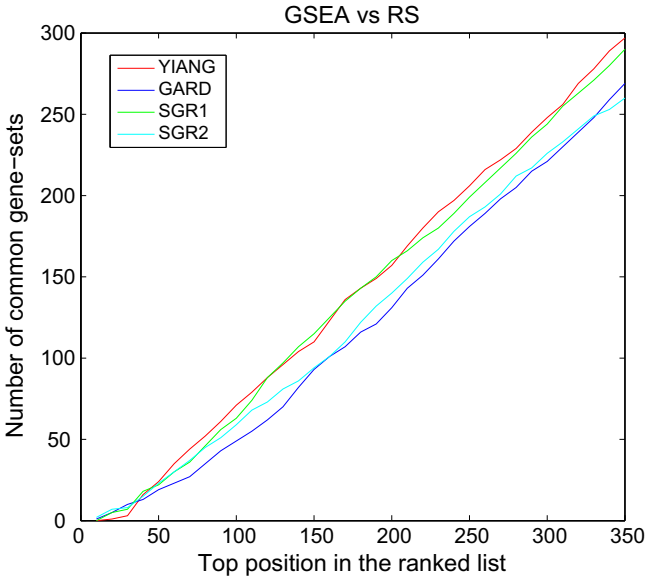

Fig. 1. Overlap of the lists of deregulated gene sets produced by GSEA and RS in the four data sets. The *x*-axis represents the size of the list and the *y*-axis represents the overlap in each pairwise comparison.

which monitor cell cycle progression ensuring an interdependency of S-phase and mitosis, the integrity of the genome and the fidelity of chromosome segregation. Cell cycle checkpoints are essential in eukaryotes for ensuring high fidelity transmission of genetic information from one generation to the next. They include DNA damage checkpoints, DNA replication checkpoints, spindle assembly checkpoints, and cytokinesis checkpoints. The first gene set analyzed was ARFPATHWAY belonging to Biocarta database. This pathway was related to cell cycle because it includes genes, such as CDKN2A, that have the ability to elicit a p53 response and a distinctive cell cycle arrest in both the G1 and G2/M phases, acting as tumor suppressors. Other gene sets of the cell cycle found deregulated in our colon cancer data sets, are BRENTANI CELL CYCLE that contains cancer related genes involved in the cell cycle [29], CELL CYCLE belonging to Gene Ontology data base, CELL CYCLE KEGG, HSA04110 CELL CYCLE [41] and GOLDRATH CELLCYCLE [34] that includes cell cycle genes induced during antigen activation of CD8+ T cells. Moreover we found CELLCYCLEPATHWAY, belonging to Biocarta data base, which analyzes the interplay of many molecules that regulate the cell cycle, underlining the key role of the cyclins which combine with cyclin dependent kinases to drive the stages of the cell cycle. A breakdown in the regulation of this cycle can lead to out of growth control and contribute to tumor formation. Defects in many of the molecules that regulate the cell cycle have been implicated in cancer. P53, the cdk inhibitors (p15, p16, p18, p19, p21, p27), and Rb are among key genes acting to keep the cell cycle from progressing until all repairs to damaged DNA have been completed. Indeed among our 52 deregulated pathways, we found P21 ANY DN and P21 EARLY DN pathways that highlight the role of p21 in p53-independent apoptosis; instead P21 P53 ANY DN, P21 P53 EARLY DN pathways consider p53-dependent p21-induced apoptosis [62] and P27PATHWAY, a Biocarta data base pathway, blocks cell cycle progression through the G1–S transition. Low levels of p27 protein were found associated with high aggressiveness and poor prognosis among patients with a variety of malignancies, including colorectal carcinoma [55]. These studies suggested that the decline in p27 levels may contribute to uncontrolled proliferation of malignant cells, because p27 is a negative regulator of the protein kinases Cdk2/cyclin E and Cdk2/cyclin A, which drive cells into the S phase of the cell division cycle [53]. The loss of p27 in patients with malignant disease results from increased protein degradation. The machinery involved in targeting p27 for degradation is an SCF type ubiquitin ligase complex that contains S phase kinase protein 2 (Skp2) as the specific substrate recognizing the subunit. Recent studies show that Skp2 is overexpressed and inversely related to p27 in colorectal carcinoma. Increased protein and gene expression of Skp2 were found to be strongly correlated with high tumor aggressiveness, suggesting that Skp2 acts as an oncogene by promoting the rate of p27 degradation [38]. Skp2 protein recruits also E2F for ubiquitination and degradation; E2F-1 is a transcription factor that regulates the expression of genes involved in the cell cycle and that is involved in progression of the cell cycle from G1 into S phase. Over-expression of E2F-1 can induce cellular transformation and its underexpression can repress apoptosis. Knowledge of E2F-1 activity and its regulation represents only a part of the complex set of interactions regulating the cell cycle and potential targets for the treatment of cancer. In our analysis we found SKP2E2FPATHWAY, belonging to Biocarta data base, that analyzes the transcription factor E2F-1 and the expression of transcriptionally regulated genes required for S-phase entry and DNA synthesis. Some findings sug-

Table 4
Number of pathways identified by GSEA and RS in three data sets simultaneously. The FP rate is reported in parentheses.

SGR1, SGR2, JIANG				SGR1, SGR2, GARD			
GSEA		RS		GSEA		RS	
UP	DOWN	UP	DOWN	UP	DOWN	UP	DOWN
97 (46%)	17 (94%)	94 (45%)	13 (92%)	76 (32%)	8 (88%)	90 (42%)	11 (91%)
SGR1, JIANG, GARD				SGR2, JIANG, GARD			
GSEA		RS		GSEA		RS	
UP	DOWN	UP	DOWN	UP	DOWN	UP	DOWN
120 (57%)	9 (89%)	110 (53%)	12 (92%)	84 (38%)	6 (83%)	100 (48%)	6 (83%)

gested that E2F-1 induces apoptotic cell death and inhibits proliferation in human colon cancer cells lines [59]. Another group of cell cycle pathways found deregulated in our analysis regards the interphase and its checkpoints that are used by the cell to monitor and regulate the progress of the cell cycle. G1PATHWAY, belonging to Biocarta data base, analyses the G1/S cell cycle checkpoint that controls the passage of eukaryotic cells from the first gap phase (G1) into the DNA synthesis phase (S). Two cell cycle kinases, CDK4/6–cyclin D and CDK2–cyclin E, and the transcription complex that includes Rb and E2F are pivotal in controlling this checkpoint. During G1 phase, the Rb-HDAC repressor complex binds to the E2F–DP1 transcription factors, inhibiting the downstream transcription. Phosphorylation of Rb by CDK4/6 and CDK2 dissociates the Rb-repressor complex, permitting transcription of S-phase genes encoding for proteins that amplify the G1 to S phase switch and that are required for DNA replication. Many different stimuli exert checkpoint control including TGFb, DNA damage, contact inhibition and growth factor withdrawal. G1 TO S CELL CYCLE REACTOME analyses the G1/S transition. In this phase the Cyclin E–Cdk2 complexes control the transition from G1 into S-phase. Finally we found G2PATHWAY, also belonging to Biocarta data base, which deals with the G2/M DNA damage checkpoint preventing the cell from entering mitosis (M phase) if the genome is damaged. The Cdc2–cyclin B kinase is pivotal in regulating this transition. Another pathway related to cell cycle was CIS XPC UP [60]. XPC is an important DNA damage recognition protein involved in DNA nucleotide excision repair. Experimental studies about the role of the XPC protein in cisplatin treatment-mediated cell cycle regulation, have shown that the cell cycle and cell proliferation-related genes were the most affected by the XPC defect in the cisplatin treatment. Many other cellular function genes were also affected by the XPC defect in the treatment; the XPC defect reduced the p53 responses to the cisplatin treatment and the ability to activate caspase-3 was also attenuated. These results suggest that the XPC protein plays a critical role in initiating the cisplatin DNA damaging treatment-mediated signal transduction process, resulting in activation of the p53 pathway and cell cycle arrest that allow DNA repair and apoptosis to take place. These results reveal an important role of the XPC protein in the cancer prevention. Also the analysis of SERUM FIBROBLAST CELLCYCLE [31] and UNDERHILL PROLIFERATION [56] showed cell cycle- and proliferation-related genes. The transcriptional signature of all these pathways could provide a powerful predictor of the clinical course in several common carcinomas, such as in colorectal cancer, predicting increased risk of metastasis and death. Besides cell cycle pathways, we found another 22 pathways overexpressed in cancer tissues and that might be good markers for prognosis of colorectal cancer. Among these we found HDACI COLON BUT12HRS DN and HDACI COLON BUT24HRS DN [44] that are two pathways obtained experimentally by SW260 colon carcinoma cells when they are downregulated by butyrate. The short-chain fatty acid butyrate is a physiological regulator of many pathways of colonic epithelial cell maturation, cell cycle arrest, lineage-specific differentiation and apoptosis. Thus microarray analysis of gene expression profiles can be used to characterize and distinguish the mechanisms of response of colonic epithelial cells to physiological and pharmacological inducers of cell maturation. This has important implications for characterization of chemopreventive agents and recognition of potential toxicity. Another two important pathways were HSA00790 FOLATE BIOSYNTHESIS and ONE CARBON POOL BY FOLATE [41] that include genes involved in folate biosynthesis. Considerable epidemiologic evidence suggests that a low-folate diet is associated with an increased risk of colorectal cancer, although the results of a recent randomized trial indicate that folate supplementation may not reduce the risk of adenoma recurrence. In laboratory models, folate deficiency appears to induce p53 mutation [49]. Furthermore we found MYC ONCOGENIC SIGNATURE [27] and SCHUMACHER MYC UP [50] pathways that consider Myc oncogene, the genes up-regulated by Myc and its role in cell cycle progression, apoptosis and cellular transformation. Mutations, overexpression, rearrangement and translocation of this gene have been associated with a variety of tumors. Myc protects from p53-mediated apoptosis. Some findings indicated that failure of the normal apoptotic process together with deregulation of Myc proto-oncogene might promote the development of colorectal tumors and its overexpression is observed in most colorectal cancers [36,51]. MANALO HYPOXIA DN [43] gene set considered genes downregulated under hypoxic conditions. Hypoxia refers to the condition that a cell experiences under oxygen deficiency. Alternatively, cancer cells can genetically elicit a hypoxic response in the setting of normal oxygen levels to activate new blood vessel formation to experience a growth advantage. For example, VEGF gene, which is generally up-regulated by hypoxic conditions, promotes normal blood vessel formation and angiogenesis related to tumor growth. Its expression is induced in colon and other cancer cells as a result of hypoxia and multiple genetic alterations; probably VEGF works as regulators of colon cancer cell invasion [37]. Two important pathways for colorectal cancer prognosis were SANSOM APC LOSS4 UP and SANSOM APC LOSS5 UP [48] that contain genes upregulated following Apc loss. Apc is well characterized as a tumor-suppressor gene in the intestine; although the function of APC, specifically mutated in familial adenomatous polyposis (FAP), is unknown, there is some evidence that it may affect apoptosis in colorectal epithelial cells. Mutations of APC gene are associated with the earliest stages of colorectal tumorigenesis [46]. Then we found other two pathways, CROONQUIST IL6 RAS DN and CROONQUIST IL6 STARVE UP [25], that show gene expression patterns involved in the effects of IL-6 response and N-ras-activating mutations. IL-6 stimulates cell growth through Ras-mitogen-activated protein kinase (MAPK) signaling pathway targets; it has been shown to be a potent mitogen and survival factor for cancer cells. In fact most upregulated genes of these gene set are involved in cell cycle progression. Also Ras protein controls cell growth and differentiation by transduction of extracellular mitogenic signals. Furthermore Ras gene mutation is an important type of somatic alteration identified in a variety of tumors including colorectal cancers; ras gene mutations may be the initiating event in colorectal tumors; adenomas with ras gene mutations may be favorable to progress [33]. DOX RESIST GASTRIC UP [42] pathway contained genes upregulated in gastric cancer cell lines resistant to doxorubicin, compared to parent chemosensitive lines. The differential expression is associated with the acquisition of resistance in human gastric cancer cells. A major obstacle in chemotherapy is treatment failure due to anticancer drug resistance. The emergence of acquired resistance results from host factors and genetic or epigenetic changes in the cancer cells. The resistance itself may be due to decreased drug accumulation, alteration of intracellular drug distribution, reduced drug–target interaction, increased detoxification response, cell-cycle deregulation, increased damaged-DNA repair and reduced apoptotic response. Many studies focus on a limited number of candidate genes in chemoresistance which will be used as novel chemotherapeutic targets for the treatment or prevention of cancer; for example, it is well known that overexpression of the multidrug resistance gene (MDR1) is associated with cancer cells that have drug resistance. Cell cycle deregulation is an important molecular event in the acquisition of drug resistance. Most of the genes identified overexpressed in doxorubicin-resistant gastric cancer cells were involved in the cell cycle. Some genes, as MDK gene, are frequently overexpressed not only in gastric cancer, but also in a variety of tumors including colon cancer. Another pathway analyzed was GAY YY1 DN [26] that include a list of YY1 target genes. YY1 transcription factor coordinates multiple

essential biological processes through a complex transcriptional network. Some findings suggested that YY1 has an important role in the control of cell growth, proliferation, apoptosis, oncogenic transformation, and differentiation. HSA03020 RNA POLYMERASE [41] pathway considers genes involved in RNA polymerase functions. Transcription of rRNA and tRNA genes by RNA polymerases I and III is essential for sustained protein synthesis and is therefore a fundamental determinant of the capacity of a cell to grow. When cell growth is not required, this transcription is repressed by retinoblastoma protein, p53 and ARF. This gene set is included in our analysis because the inactivation of these tumor suppressors in cancers deregulates RNA polymerases I and III, and oncoproteins, such as Myc, can stimulate these systems further. Such events might have a significant impact on the growth potential of tumors [61]. In the comparison of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine) vs. fibroblasts treated with IDX + TSA (trichostatin A) to prevent differentiation, IDX TSA UP CLUSTER3 pathway [30] consists of strongly up-regulated genes. TSA is an inhibitor of histone deacetylases (HDAC). HDACs are generally associated with gene repression. Compared to non-malignant cells, colon cancer cells exhibit increased HDACs activity. HDAC1 and HDAC3 are upregulated in colon cancer cells and in primary colon cancer; moreover silencing of HDAC1 and HDAC3 in colon cancer cells induces apoptosis [54]. Then we analyzed INOS ALL UP [63] pathway that encloses several modulated families of genes, including genes coding for proinflammatory transcription factors, cytokines, cytokine receptors, proteins associated with cell proliferation and cellular energetics, as well as proteins involved in apoptosis. In this study it was seen that inducible nitric oxide synthase (iNOS) acts to suppress proliferation and protein synthesis and modulates several genes associated with apoptosis, leading to a full anti-apoptotic effect. IRITANI ADPROX UP [39] pathway analysed the biological balance between Myc and Mad gene levels that control expression of growth regulating genes. MAD proteins antagonize the functions of MYC oncoproteins, and the latter are deregulated in the majority of human cancers. While MYC sensitizes cells to pro-apoptotic signals, the transcriptional repressor MAD1 inhibits apoptosis in response to a broad range of stimuli, including oncoproteins. PEART HISTONE DN [45] gene set included genes related to cell proliferation downregulated by SAHA and depsipeptide, which are histone deacetylase inhibitors (HDACis). May be that, through the ability of HDACis of regulating the expression of specific proliferative and/or apoptotic genes, growth and survival of tumor cell are inhibited. They regulate the expression of several genes within distinct apoptosis and cell cycle pathways. Then we analyzed SHEPARD GENES COMMON BW CB MO [52] pathway which includes genes associated with cell cycle and cancer susceptibility. One gene identified in this study is B-myb, which is part of a small family of transcription factors with characteristic regions of homology that includes the c-myb proto-oncogene. Some studies showed that B-myb plays a role in cell cycle regulation, particularly at the G1/S transition, and that loss of function of this gene is associated with cancer. Another two pathways found deregulated in our analysis were VANTVEER BREAST OUTCOME GOOD VS POOR DN [57] that consists of genes regulating cell cycle, invasion, metastasis and angiogenesis, and VERNELL PRB CLSTR1 [58] that contains pRB pathway target genes. This last pathway included the genes found down-regulated by pRB and p16 and up-regulated by E2F. Deregulation of the retinoblastoma protein (pRB) pathway is a hallmark of human cancer [32]. The core members of this pathway include the tumor suppressor protein pRB which regulates progression through the cell division cycle. The expression of pRB and p16 resulted in significant repression and activation of a large number of genes. Transcriptional changes were found in genes that are essential for DNA replication and cell proliferation. Then we found two pathways, CANCER NEOPLASTIC META UP and CANCER UNDIFFERENTIATED META UP [47], that included genes upregulated in cancer vs normal tissues comparison in the first pathway, and in undifferentiated vs well-differentiated tumors comparison in the second gene set. These two gene sets highlight a transcriptional profile that is commonly activated in various types of undifferentiated cancer: this suggests common molecular mechanisms by which cancer cells progress and avoid differentiation. Another two interesting pathways found deregulated in our analysis were BRCA PROGNOSIS NEG [57] that contains genes whose expression is negatively correlated with breast cancer outcomes and BREAST DUCTAL CARCINOMA GENES that includes genes upregulated in breast tumors. Likely we found these pathway because some findings suggested an association between the risk of breast and colorectal cancers. Some evidences highlight an important role for insulin and insulin-like growth factors (IGFs) in the promotion of carcinogenesis in both organs. Also BRCA1 gene that acts as a tumor suppressor, in some studies was found deregulated in the two types of cancer. In fact defects in BRCA1 are a cause of genetic susceptibility to breast cancer and mutations in BRCA1 are thought to be responsible for 45% of inherited breast cancer; however BRCA1 mutation carriers have a 4-fold increased risk of colon cancer. Loss of heterozygosity at the BRCA1 gene locus was shown to be associated with shorter survival in colorectal cancer; moreover recent evidences showed that the expression of ATM and BRCA1 is a prognostic marker in colorectal cancer [35]. The only down-regulated pathway identified (1 AND 2 METHYLNAPHTHALENE DEGRADATION) includes genes belonging to the alcohol dehydrogenase family. Members of this enzyme family metabolize a wide variety of substrates, including ethanol, retinol, other aliphatic alcohols, hydroxysteroids, and lipid peroxidation products. The enzyme encoded by ADH7 gene is active as a retinol dehydrogenase; thus it may participate in the synthesis of retinoic acid, a hormone important for cellular differentiation.

5. Discussion and conclusions

In this paper, we have addressed the problem of reproducibility of results of pathway analysis in genome-wide expression studies. In particular we have assessed if lists of pathways obtained in different experiments, in which the same phenotypic conditions were assayed, shown a statistically significant and biologically relevant overlap. To this end, we have used MSigDB [9] a vast collections of 1891 curated gene sets coding biological processes, cellular functions and in general gene networks defined experimentally as well as on the basis of a-priori knowledge. The deregulation of pathways was assessed through GSEA [9] and RS [20], two methods which implement different statistical schemes for measuring association of groups of genes to the phenotype. Finally, we analyzed the results obtained by these two methods applied to three different gene expression data sets [17–19] plus a new data set collected in Casa Sollievo della Sofferenza Hospital, Foggia, Italy, relative to subjects affected by colorectal cancer (see Section 2).

The main conclusion we can draw on the basis of our extensive statistical assessment is that the results of pathway analysis are highly reproducible. We determined a set composed of 53 pathways, simultaneously altered in all the four data sets analyzed, having strong biological implications with the pathology, which provides a signature of colorectal cancer in terms of deregulated gene networks. They were determined independently by GSEA and RS, two deeply different methods which aim at evaluating the association of pathways with phenotypes. The former uses a simple phenotypic permutation for evaluating the significance of the enrichment of a signature. The latter couples the classic permutation scheme to a novel restandardization procedure which aims

Table 5
Statistical significance and FDR of the set of 53 deregulated pathways as evaluated by GSEA in the four data sets. The first 52 pathways are up-regulated and the last is the only down-regulated gene set.

Pathway	JIANG		GARD		SGR1		SGR2		Rank
	P	FDR	*P*	FDR	*P*	FDR	*P*	FDR	
SANSOM APC LOSS4 UP	0.0E+00	0.0%	7.5E−03	50.2%	0.0E+00	4.9%	2.1E−03	100.0%	8
BRCA PROGNOSIS NEG	0.0E+00	0.1%	2.0E−03	27.1%	0.0E+00	2.1%	3.9E−02	43.7%	11
HSA04110 CELL CYCLE	0.0E+00	0.1%	3.9E−03	48.7%	0.0E+00	2.1%	0.0E+00	100.0%	14
CELL CYCLE	0.0E+00	0.1%	2.4E−02	27.0%	0.0E+00	2.0%	4.1E−03	52.4%	18
G1 TO S CELL CYCLE REACTOME	0.0E+00	0.1%	4.8E−02	29.5%	0.0E+00	2.4%	1.2E−02	62.4%	24
SANSOM APC LOSS5 UP	0.0E+00	0.1%	0.0E+00	36.7%	0.0E+00	2.2%	1.7E−02	57.8%	29
BRENTANI CELL CYCLE	0.0E+00	0.1%	1.8E−02	35.0%	0.0E+00	2.3%	8.0E−03	52.5%	29
CELL CYCLE KEGG	0.0E+00	0.1%	1.8E−02	26.4%	0.0E+00	2.2%	1.0E−02	57.4%	32
LI FETAL VS WT KIDNEY DN	0.0E+00	0.1%	9.9E−03	26.3%	0.0E+00	2.0%	1.6E−02	53.7%	32
GOLDRATH CELL CYCLE	0.0E+00	0.1%	3.9E−03	34.9%	0.0E+00	3.0%	1.6E−02	40.8%	38
BROWN MYELOID PROLIF AND SELF RENEWAL	0.0E+00	0.1%	0.0E+00	28.6%	2.0E−03	2.8%	1.8E−02	57.9%	40
CANCER NEOPLASTIC META UP	0.0E+00	0.2%	6.0E−03	37.9%	1.9E−03	2.4%	2.0E−03	59.9%	42
UNDERHILL PROLIFERATION	0.0E+00	0.1%	9.8E−03	24.2%	0.0E+00	2.2%	2.4E−02	39.9%	44
CELLCYCLEPATHWAY	0.0E+00	0.1%	1.6E−02	34.2%	0.0E+00	2.4%	2.3E−02	51.6%	50
VANTVEER BREAST OUTCOME GOOD VS POOR DN	0.0E+00	0.2%	7.9E−03	24.6%	0.0E+00	3.3%	2.9E−02	45.4%	51
GAY YY1 DN	0.0E+00	0.1%	1.2E−02	27.0%	0.0E+00	2.3%	4.8E−02	39.6%	53
HDACI COLON BUT12HRS DN	2.1E−03	0.4%	6.1E−03	25.7%	2.0E−03	3.5%	1.4E−02	62.2%	54
HDACI COLON BUT24HRS DN	6.1E−03	1.4%	8.1E−03	27.4%	2.0E−03	5.7%	1.0E−02	52.3%	55
ADIP DIFF CLUSTER5	0.0E+00	0.4%	4.3E−02	26.8%	0.0E+00	2.4%	6.2E−03	50.5%	56
IRITANI ADPROX UP	0.0E+00	0.1%	1.2E−02	27.5%	0.0E+00	4.4%	2.3E−02	53.5%	56
P21 ANY DN	0.0E+00	0.1%	1.3E−02	24.0%	2.0E−03	6.2%	6.0E−03	52.5%	56
SHEPARD GENES COMMON BW CB MO	0.0E+00	0.0%	9.7E−03	81.0%	2.0E−03	2.5%	2.1E−02	68.8%	60
CANCER UNDIFFERENTIATED META UP	0.0E+00	0.4%	3.9E−03	43.9%	8.0E−03	6.4%	2.2E−02	46.9%	63
ADIP DIFF CLUSTER4	0.0E+00	0.3%	2.7E−02	27.3%	1.9E−03	4.6%	2.6E−02	46.6%	66
LEE TCELLS3 UP	0.0E+00	0.1%	8.0E−03	46.7%	4.1E−03	3.1%	3.0E−02	51.6%	70
SERUM FIBROBLAST CELLCYCLE	0.0E+00	0.1%	1.2E−02	26.2%	2.0E−03	2.4%	2.6E−02	41.8%	71
PEART HISTONE DN	0.0E+00	0.1%	0.0E+00	27.2%	4.1E−03	3.2%	4.6E−02	50.7%	72
MYC ONCOGENIC SIGNATURE	0.0E+00	0.4%	2.9E−02	27.2%	2.0E−03	3.0%	2.3E−02	100.0%	72
OLDAGE DN	0.0E+00	0.1%	2.6E−02	27.0%	4.0E−03	3.3%	2.0E−02	39.8%	73
CIS XPC UP	0.0E+00	0.3%	2.0E−02	27.9%	2.0E−03	3.1%	2.8E−02	50.2%	75
P21 P53 ANY DN	0.0E+00	0.2%	1.6E−02	24.4%	2.0E−03	3.1%	3.8E−02	40.0%	79
INOS ALL UP	0.0E+00	0.4%	1.9E−02	27.3%	3.9E−03	3.0%	1.9E−02	53.1%	81
ONE CARBON POOL BY FOLATE	0.0E+00	0.4%	4.5E−02	28.1%	0.0E+00	2.1%	2.9E−02	66.8%	81
SCHUMACHER MYC UP	0.0E+00	0.1%	2.3E−02	29.0%	6.0E−03	3.9%	1.9E−02	53.2%	83
G1PATHWAY	2.0E−03	0.5%	4.0E−02	27.0%	0.0E+00	2.1%	6.3E−03	60.7%	88
MANALO HYPOXIA DN	0.0E+00	0.2%	1.6E−02	25.0%	3.9E−03	2.4%	3.6E−02	44.3%	88
CROONQUIST IL6 STARVE UP	2.0E−03	1.2%	1.8E−02	25.5%	0.0E+00	5.9%	2.6E−02	40.0%	89
IDX TSA UP CLUSTER3	0.0E+00	0.1%	1.6E−02	23.7%	4.0E−03	3.1%	3.4E−02	43.8%	89
ZHAN MM CD138 PR VS REST	0.0E+00	0.7%	0.0E+00	28.0%	8.0E−03	6.6%	3.0E−02	40.6%	90
SHIPP FL VS DLBCL DN	0.0E+00	0.1%	2.9E−02	27.4%	1.2E−02	4.3%	4.0E−03	55.2%	93
VERNELL PRB CLSTR1	0.0E+00	0.1%	3.2E−02	27.0%	4.0E−03	4.6%	3.5E−02	43.8%	97
G2PATHWAY	2.0E−03	0.2%	4.9E−02	26.8%	0.0E+00	2.0%	2.7E−02	52.2%	110
CROONQUIST IL6 RAS DN	0.0E+00	1.1%	1.9E−02	25.2%	6.0E−03	7.7%	4.3E−02	41.3%	111
P21 P53 EARLY DN	2.1E−03	1.8%	1.4E−02	28.3%	1.2E−02	10.7%	1.2E−02	39.3%	119
P27PATHWAY	4.1E−03	1.6%	9.6E−03	24.9%	1.4E−02	7.2%	2.7E−02	45.5%	121
ARFPATHWAY	2.0E−03	1.8%	2.4E−02	26.2%	6.0E−03	6.0%	1.9E−02	59.1%	123
HSA03020 RNA POLYMERASE	2.1E−03	3.5%	5.0E−02	27.1%	0.0E+00	2.2%	3.1E−02	47.5%	126
P21 EARLY DN	0.0E+00	1.3%	3.7E−02	26.9%	9.9E−03	5.3%	0.0E+00	52.1%	127
DOX RESIST GASTRIC UP	0.0E+00	0.4%	3.9E−02	27.7%	6.0E−03	6.0%	4.9E−02	41.0%	127
BREAST DUCTAL CARCINOMA GENES	2.1E−03	1.1%	3.9E−02	27.5%	4.0E−03	4.3%	1.0E−02	42.1%	128
HSA00790 FOLATE BIOSYNTHESIS	2.0E−03	1.3%	3.6E−02	27.3%	7.9E−03	3.3%	2.1E−03	100.0%	139
SKP2E2FPATHWAY	3.6E−02	5.6%	2.0E−02	27.0%	2.6E−02	8.8%	3.2E−02	41.2%	154
1 AND 2 METHYLNAPHTHALENE DEGRADATION	2.9E−02	17.0%	0.0E+00	59.7%	1.4E−02	20.4%	3.8E−02	62.5%	63

at evaluating how the deregulation depends on the identity of the genes present in the pathway. In particular, it aims at assessing if lists of the same size composed of randomly selected genes from the ones present on the microarray produce comparable enrichments. In fact as admirably pointed out in [21] "any method for assessing gene-sets should compare a given gene-set score not only to scores from permutations of the sample labels, but also has to take into account scores from sets formed by random selections of genes".

Our study reveals that, although reproducible, the results of pathway analysis have to be interpreted with caution when limited to a single data set and, more importantly, suggests alternative and more robust procedures of analysis of expression data. In fact, considering gene sets enriched in tumor samples only (see Table 2), both methods have a FP rate in the range [65%, 85%]. This means that more than half of the pathways identified as deregulated by a method when applied to a single data set, are not confirmed by other studies analyzing the same phenotype. To reduce the FP rate in the analysis of pathways we suggest an alternative approach consisting in intersecting lists of deregulated pathways identified in different data sets. In fact, as Table 3 shows, the FP rate decreases to [45%, 75%] and [50%, 65%] for GSEA and RS, respectively, if we intersect lists of deregulated pathways identified in two different data sets. The FP rate reduces further in the ranges [32%,

Table 6
Statistical significance and FDR of the set of 53 deregulated pathways as evaluated by RS in the four data sets. The first 52 pathways are up-regulated and the last is the only down-regulated gene set.

Pathway	JIANG		GARD		SGR1		SGR2		Rank
	P	FDR	*P*	FDR	*P*	FDR	*P*	FDR	
P21 P53 ANY DN	0.0E+00	0.0%	0.0E+00	0.0%	0.0E+00	0.0%	9.0E−03	23.4%	8
OLDAGE DN	0.0E+00	0.0%	0.0E+00	0.0%	0.0E+00	0.0%	3.0E−03	20.9%	13
ZHAN MM CD138 PR VS REST	0.0E+00	0.0%	0.0E+00	0.0%	1.0E−03	3.5%	1.7E−02	26.5%	13
CANCER NEOPLASTIC META UP	0.0E+00	0.0%	0.0E+00	0.0%	0.0E+00	0.0%	4.0E−03	22.9%	14
CROONQUIST IL6 RAS DN	0.0E+00	0.0%	1.0E−02	12.5%	0.0E+00	0.0%	6.0E−03	23.4%	17
P21 P53 EARLY DN	0.0E+00	0.0%	0.0E+00	0.0%	1.0E−03	3.5%	2.2E−02	29.4%	17
CANCER UNDIFFERENTIATED META UP	0.0E+00	0.0%	1.0E−03	3.6%	0.0E+00	0.0%	1.2E−02	24.8%	20
LI FETAL VS WT KIDNEY DN	0.0E+00	0.0%	1.0E−03	3.6%	0.0E+00	0.0%	1.0E−03	12.8%	20
ADIP DIFF CLUSTER5	0.0E+00	0.0%	1.5E−02	15.5%	0.0E+00	0.0%	6.0E−03	23.4%	22
IDX TSA UP CLUSTER3	0.0E+00	0.0%	0.0E+00	0.0%	1.0E−03	3.5%	1.6E−02	25.7%	23
SERUM FIBROBLAST CELL CYCLE	0.0E+00	0.0%	0.0E+00	0.0%	1.0E−03	3.5%	1.2E−02	24.8%	23
BRCA PROGNOSIS NEG	2.0E−03	4.2%	0.0E+00	0.0%	1.0E−03	3.5%	3.0E−03	20.9%	26
HSA04110 CELL CYCLE	0.0E+00	0.0%	0.0E+00	0.0%	0.0E+00	0.0%	1.4E−02	25.0%	27
SANSOM APC LOSS4 UP	0.0E+00	0.0%	1.1E−02	13.4%	0.0E+00	0.0%	1.0E−03	12.8%	29
VANTVEER BREAST OUTCOME GOOD VS POOR DN	1.0E−03	2.7%	2.0E−03	5.4%	0.0E+00	0.0%	0.0E+00	0.0%	30
CELL CYCLE	0.0E+00	0.0%	1.0E−03	3.6%	1.0E−03	3.5%	1.0E−02	24.2%	31
DOX RESIST GASTRIC UP	0.0E+00	0.0%	1.0E−03	3.6%	2.0E−03	5.6%	1.9E−02	27.9%	32
CELL CYCLE KEGG	0.0E+00	0.0%	1.0E−03	3.6%	1.0E−03	3.5%	1.4E−02	25.0%	34
LEE TCELLS3 UP	0.0E+00	0.0%	0.0E+00	0.0%	2.0E−03	5.6%	2.1E−02	28.7%	34
SHIPP FL VS DLBCL DN	1.0E−03	2.7%	0.0E+00	0.0%	1.8E−02	18.9%	3.0E−03	20.9%	35
MANALO HYPOXIA DN	0.0E+00	0.0%	0.0E+00	0.0%	3.0E−03	7.1%	2.1E−02	28.7%	36
BRENTANI CELL CYCLE	0.0E+00	0.0%	1.0E−03	3.6%	0.0E+00	0.0%	1.1E−02	24.2%	37
UNDERHILL PROLIFERATION	0.0E+00	0.0%	4.0E−03	7.7%	0.0E+00	0.0%	8.0E−03	23.4%	37
GAY YY1 DN	6.0E−03	8.2%	1.0E−03	3.6%	0.0E+00	0.0%	6.0E−03	23.4%	38
CROONQUIST IL6 STARVE UP	0.0E+00	0.0%	5.0E−03	8.3%	1.0E−03	3.5%	1.1E−02	24.2%	40
P21 ANY DN	0.0E+00	0.0%	2.0E−03	5.4%	5.0E−03	9.8%	7.0E−03	23.4%	41
GOLDRATH CELL CYCLE	0.0E+00	0.0%	1.0E−03	3.6%	4.0E−03	8.2%	1.1E−02	24.2%	42
HDACI COLON BUT12HRS DN	0.0E+00	0.0%	1.0E−03	3.6%	1.0E−02	14.4%	8.0E−03	23.4%	44
SANSOM APC LOSS5 UP	3.0E−03	5.4%	0.0E+00	0.0%	9.0E−03	13.6%	0.0E+00	0.0%	47
G2PATHWAY	0.0E+00	0.0%	3.0E−03	6.5%	2.0E−03	5.6%	1.4E−02	25.0%	55
P21 EARLY DN	3.0E−03	5.4%	1.7E−02	16.4%	2.0E−03	5.6%	2.0E−03	17.0%	57
BROWN MYELOID PROLIF AND SELF RENEWAL	1.0E−03	2.7%	3.0E−03	6.5%	4.0E−03	8.2%	6.0E−03	23.4%	58
VERNELL PRB CLSTR1	0.0E+00	0.0%	1.0E−03	3.6%	1.1E−02	14.8%	2.3E−02	29.6%	58
ADIP DIFF CLUSTER4	1.0E−03	2.7%	5.0E−03	8.3%	0.0E+00	0.0%	1.3E−02	25.0%	59
HSA00790 FOLATE BIOSYNTHESIS	2.0E−03	4.2%	2.0E−03	5.4%	3.3E−02	26.1%	2.0E−03	17.0%	59
SCHUMACHER MYC UP	2.0E−03	4.2%	3.0E−03	6.5%	1.3E−02	16.5%	9.0E−03	23.4%	59
BREAST DUCTAL CARCINOMA GENES	0.0E+00	0.0%	4.0E−03	7.7%	6.0E−03	10.9%	1.1E−02	24.2%	62
G1PATHWAY	4.0E−03	6.5%	1.0E−02	12.5%	0.0E+00	0.0%	7.0E−03	23.4%	62
CIS XPC UP	1.0E−03	2.7%	3.0E−03	6.5%	7.0E−03	11.7%	1.3E−02	25.0%	70
G1 TO S CELL CYCLE REACTOME	0.0E+00	0.0%	8.0E−03	11.3%	2.0E−03	5.6%	4.5E−02	37.9%	74
ARFPATHWAY	2.6E−02	19.6%	3.0E−03	6.5%	1.1E−02	14.8%	8.0E−03	23.4%	75
IRITANI ADPROX UP	3.0E−03	5.4%	4.0E−03	7.7%	1.2E−02	15.5%	1.4E−02	25.0%	77
HDACI COLON BUT24HRS DN	8.0E−03	9.6%	2.0E−03	5.4%	4.6E−02	30.6%	5.0E−03	23.1%	84
PEART HISTONE DN	4.0E−03	6.5%	1.0E−03	3.6%	1.5E−02	17.6%	1.7E−02	26.5%	87
SHEPARD GENES COMMON BW CB MO	0.0E+00	0.0%	2.8E−02	21.2%	0.0E+00	0.0%	3.7E−02	35.4%	90
MYC ONCOGENIC SIGNATURE	2.0E−03	4.2%	1.7E−02	16.4%	1.7E−02	18.1%	3.0E−03	20.9%	91
ONE CARBON POOL BY FOLATE	6.0E−03	8.2%	7.0E−03	10.6%	1.0E−03	3.5%	3.3E−02	33.3%	93
INOS ALL UP	7.0E−03	8.9%	5.0E−03	8.3%	1.9E−02	19.5%	2.2E−02	29.4%	102
CELLCYCLEPATHWAY	1.1E−02	11.7%	1.5E−02	15.5%	0.0E+00	0.0%	1.8E−02	27.1%	106
SKP2E2FPATHWAY	2.1E−02	17.1%	5.0E−03	8.3%	3.5E−02	27.1%	6.0E−03	23.4%	111
HSA03020 RNA POLYMERASE	1.7E−02	15.1%	5.0E−03	8.3%	1.2E−02	15.5%	4.4E−02	37.6%	116
P27PATHWAY	1.8E−02	15.7%	1.5E−02	15.5%	4.2E−02	29.4%	3.2E−02	33.2%	138
1 AND 2 METHYLNAPHTHALENE DEGRADATION	1.5E−02	14.0%	1.5E−02	15.5%	1.0E−03	3.5%	7.0E−03	23.4%	15

57%] and [42%, 53%] for GSEA and RS, respectively, if we intersect lists of pathways identified in three different data sets (see Table 4). Alternatively, when different data sets relative to the same phenotypic conditions are not available, our suggestion is to intersect lists of deregulated pathways identified by different methods applied on the same data set.

The proposed approach of intersecting lists of pathways identified in different data sets highlights an important aspect concerning the usual estimators adopted for quantifying the statistical significance of discoveries. In statistical hypothesis testing, the *P*-value *P* is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. This indicator is usually used in the case of single hypothesis testing. In the case of multiple hypothesis testing, the simple *P*-value is not useful anymore as indicator of the statistical significance. In fact, if we test n hypotheses in which the null hypothesis is true, then by chance on average we will reject erroneously $n\alpha$ null hypotheses because $P \leqslant \alpha$, with α as significance level of the test. In the case of multiple hypothesis testing a more suitable statistical indicator is the false discovery rate (FDR) which takes into account the number of hypotheses simultaneously tested. In fact it represents the rate of false discoveries among the s hypotheses accepted as significant: $FDR = n\alpha/s$. So, if we accept s hypotheses as statistically significant with an FDR of 25%, then it means that the 25% of s are false discoveries. Also the FDR is not free of faults. In fact, the FDR is not a good statistical indicator in those experimental conditions, as the ones considered in this work, in which the ratio between the number of true alternative hypotheses (the size of the

gold standard) and the number of hypothesis tested (the number of pathways in MSigDB) is close to zero. At 0.05 level, the number of pathways associated to the phenotype by chance (95) is larger than the number of pathways really associated. In this experimental conditions the signal is overwhelmed in the noise and it is difficult to separate the signal from the noise. As Tables 5 and 6 show, the FDR is not a reliable indicator of the statistical significance of pathways associated to the phenotype. In fact many pathways belonging to the gold standard have FDR greater than the 30% in two data sets and would have not been taken into account in successive analysis. If a gene set has low *P*-value and high FDR in a single data set, we cannot consider it as significant. We can consider it as significant if the gene set receives low *P*-values in different data sets. This is particularly important when the association between gene set and phenotypical conditions is weak, that is when the involvement of the pathway in the pathology is modest. The methodology we propose is able to identify also weak association signals comparing results obtained in different data sets.

Our genome-wide expression analysis of pathways altered in subjects affected by colorectal cancer has highlighted numerous markers well known to be associated to the pathology at hand. The same approach may be pursued for understanding the genetics basis of other complex and polygenic diseases and for dissecting the molecular mechanisms altered in onset and progression of neoplasia.

Conflict of interest

None declared.

Funding

This work was supported by grants from Regione Puglia, Progetto Strategico PS_012 and Progetto Reti di Laboratori Pubblici di Ricerca BISIMANE.

Authors contributions

N.A. and S.M. conceived the study. R.M., and A.D'.A designed the algorithms and A.P., M.C. and O.P. conduced the experiments and, together with ADi evaluated and compared the experimental results. All the authors contributed to the drafting of the article.

Acknowledgement

A.D'.A. is a PhD student of Dipartimento Interateneo di Fisica, Univerisitá degli Studi di Bari, Italy. O.P. is a PhD student of Dipartimento di Biochimica e Biologia Molecolare "E. Quagliariello", Università degli Studi di Bari, Italy.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2009.09.005.

References

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[2] Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
[3] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. Science 1995;270:467–70.
[4] Manoli T, Gretz N, Grone HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics 2006;22:25002506.
[5] Michiels S, Koscielny S, Hill C. Predictor of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.
[6] Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
[7] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.
[8] Hsieh WP, Chub TM, Wolfinger RD, Gibson G. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics 2003;165:747–57.
[9] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[10] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 2005;102:13544–9.
[11] Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
[12] Creighton CJ. Multiple oncogeneic pathway signatures shows coordinate expression patterns in human prostate tumors. PLoS ONE 2008;3(3):e1816.
[13] Maglietta R, Piepoli A, Catalano D, Licciulli F, Carella M, Liuni S, et al. Statistical assessment of functional categories of genes deregulated in pathological conditions by using microarray data. Bioinformatics 2007;23(16):20632072.
[14] Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC. Getting started in biological pathway construction and analisys. PLoS Comput Biol 2008;4(2):e16.
[15] Papoulis A. The Fourier integral and its applications. McGraw-Hill; 1962.
[16] Girosi F, Jones M, Poggio T. Regularization theory and neural networks architectures. Neural Comput 1995;7(2):219–69.
[17] Ancona N, Maglietta R, Piepoli A, D'Addabbo A, Cotugno R, Savino M, et al. On the statistical assessment of classifiers using DNA microarray data. BMC Bioinform 2006;19:7:387.
[18] Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 2006;7:325.
[19] Jiang X, Tan J, Li J, Kivimäe S, et al. DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. Cancer Cell 2008;13(6):529–41.
[20] Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat 2007;1(1):85106.
[21] Efron B, Tibshirani R. On thesting the significance of sets of genes. Ann Appl Stat 2007;1(1):107129.
[22] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
[23] Good P. Permutation tests: a practical guide to resampling methods for testing hypothesis. Springer; 1994.
[24] Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21:1943–9.
[25] Croonquist PA, Linden MA, Zhao F, Van Ness BG. Gene profiling of a myeloma cell line reveals similarities and unique signatures among il-6 response, n-ras-activating mutations, and coculture with bone marrow stromal cells. Blood 2003;102(7):2581–92.
[26] Affar EB, Gay F, Shi Y, Liu H, Huarte M, Wu S, et al. Essential dosage-dependent functions of the transcription factor yin yang 1 in late embryonic development and cell cycle progression. Mol Cell Biol 2006;26(9):3565–81.
[27] Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439(7074):353–7.
[28] Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, et al. Gene expression in colorectal cancer 1,2. Cancer Res 2002;62:4352–63.
[29] Brentani H, Caballero OL, Camargo AA, da Silva AM, Araujo da Silva Jr W, Neto ED, et al. The human cancer genome project/cancer genome anatomy project annotation consortium, and the human cancer genome project sequencing consortium, the generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. PNAS 2003;100(23):13418–23.
[30] Burton GR, Nagarajan R, Peterson CA, McGehee Jr RE. Microarray analysis of differentiation-specific gene expression during 3t3-l1 adipogenesis. Gene 2004;329:167–85.
[31] Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, et al. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. PLoS Biol 2004;2(2):206–14.
[32] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100(1):57–70.
[33] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759–67.
[34] Goldrath AW, Luckey CJ, Park R, Benoist C, Mathis D. The molecular program induced in t cells undergoing homeostatic proliferation. Proc Natl Acad Sci USA 2004;101:16885–90.
[35] Grabsch H, Dattani M, Barker L, Maughan N, Maude K, Hansen O, et al. Expression of dna double-strand break repair proteins atm and brca1 predicts survival in colorectal cancer. Clin Cancer Res 2006;12(5):1494–500.

[36] Greco C, Alvino S, Buglioni S, Assisi D, Lapenta R, Grassi A, et al. Anticancer Res 2001;21(5):3185–92.
[37] Han J, Xia C, Gao J, Xing C, Yang X, Tang X, et al. xpression of vascular endothelial growth factor in colorectal cancer and its clinical significance. Zhonghua Yi Xue Za Zhi 2002;82(7):481–3.
[38] Hershko D, Bornstein G, Ben-Izhak O, Carrano A, Pagano M, Krausz MM, et al. Inverse relation between levels of p27(kip1) and of its ubiquitin ligase subunit skp2 in colorectal carcinomas. Cancer 2001;91:1745–51.
[39] Iritani BM, Delrow J, Grandori C, Gomez I, Klacking M, Carlos LS, et al. growth and cell size by the myc antagonist and transcriptional repressor mad1. EMBO J 2002;21(18):4820–30.
[40] Irizarry RA, Zhijin W, Harris AJ. Comparison of affymetrix genechip expression measures. Bioinformatics 2006;22:789–94.
[41] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. Kegg for linking genomes to life and the environment. Nucleic Acids Res 2008;36:D480–4.
[42] Kang HC, Kim I, Park JH, Shin Y, Ku JL, Jung MS, et al. Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. Clin Cancer Res 2004;10:272–84.
[43] Manalo DJ, Rowan A, Lavoie T, Natarajan L, Kelly BD, Ye SQ, et al. Transcriptional regulation of vascular endothelial cell responses to hypoxia by hif-1. Blood 2005;105(2):659–69.
[44] Mariadason JM, Corner GA, Augenlicht LH. Genetic reprogramming in pathways of colonic cell maturation induced by short chain fatty acids: comparison with trichostatin a, sulindac, and curcumin and implications for chemoprevention of colon cancer. Cancer Res 2000;60(16):4561–72.
[45] Peart MJ, Smyth GK, van Laar RK, Bowtell DD, Richon VM, Marks PA, et al. Identification and functional significance of genes regulated by structurally different histone deacetylase inhibitors. Proc Natl Acad Sci USA 2005;102(10): 3697–702.
[46] Morin PJ, Vogelstein B, Kinzler KW. Apoptosis and apc in colorectal tumorigenesis. Proc Natl Acad Sci USA 1996;93:7950–4.
[47] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 2004;101(25):9309–14.
[48] Sansom OJ, Reed KR, Hayes AJ, Ireland H, Brinkmann H, Newton IP, et al. Loss of apc in vivo immediately perturbs wnt signaling, differentiation, and migration. Genes Dev 2004;18(12):1385–90.
[49] Schernhammer ES, Ogino S, Fuchs CS. Folate and vitamin b6 intake and risk of colon cancer in relation to p53 expression. Gastroenterology 2008;135:770–80.
[50] Schuhmacher M, Kohlhuber F, Hlzel M, Kaiser C, Burtscher H, Jarsch M, et al. The transcriptional program of a human b cell line in response to myc. Nucleic Acids Res 2001;29(2):397–406.
[51] Seidler HBK, Utsuyama M, Nagaoka S, Takemura T, Kitagawa M, Hirokawa K. Expression level of wnt signaling components possibly influences the biological behavior of colorectal cancer in different age groups. Exp Mol Pathol 2004(76):224–33.
[52] Shepard JL, Amatruda JF, Stern HM, Subramanian A, Finkelstein D, Ziai J, et al. A zebrafish bmyb mutation causes genome instability and increased cancer susceptibility. Proc Natl Acad Sci USA 2005;102(37):13194–9.
[53] Sherr CJ, Roberts JM. Cdk inhibitors: positive and negative regulators of g1-phase progression. Genes Dev 1999;13:1501–12.
[54] Thangaraju M, Carswell KN, Prasad PD, Ganapathy V. Colon cancer cells maintain low levels of pyruvate to avoid cell death caused by inhibition of hdac1/hdac3. Biochem J 2008.
[55] Tsihlias J, Kapusta L, Slingerland J. The prognostic significance of altered cyclin dependent kinase inhibitors in human cancer. Annu Rev Med 1999; 50:401–23.
[56] Underhill GH, George D, Bremer EG, Kansas GS. Gene expression profiling reveals a highly specialized genetic program of plasma cells. Blood 2003;101(10):4013–21.
[57] van't Veer LJ, Dai H H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–6.
[58] Vernell R, Helin K, Muller H. Identification of target genes of the p16ink4a-prb-e2f pathway. J Biol Chem 2003;278(46):46124–37.
[59] Vorburger SA, Pataer A, Yoshida K, Liu Y, Lu X, Swisher SG, et al. The mitochondrial apoptosis-inducing factor plays a role in e2f-1induced apoptosis in human colon cancer cells. Ann Surg Oncol 2003;10(3):314322.
[60] Wang G, Chuang L, Zhang X, Colton S, Dombkowski A, Reiners J, et al. The initiative role of xpc protein in cisplatin dna damaging treatment-mediated cell cycle regulation. Nucleic Acids Res 2004;32:2231–40.
[61] White RJ. Rna polymerases i and iii, growth control and cancer. Nat Rev Mol Cell Biol 2005;6:67–9.
[62] Wu Q, Kirschmeier P, Hockenberry T, Yang TY, Brassard DL, Wang L, et al. Transcriptional regulation during p21waf1/cip1-induced apoptosis in human ovarian cancer cells. J Biol Chem 2002;277(39):36329–37.
[63] Zamora R, Vodovotz Y, Aulak KS, Kim PK, Kane JM, Alarcon L, et al. A dna microarray study of nitric oxide-induced genes in mouse hepatocytes: implications for hepatic heme oxygenase-1 expression in ischemia/reperfusion. Nitric Oxide 2002;7(3):165–86.